Hidden Vulnerabilities and Licensing Risks in LLM Pre-Training Datasets

Mahmoud Jahanshahi

THE UNIVERSITY OF TENNESSEE KNOXVILLE

June 2025

Published Research

- M. Jahanshahi and A. Mockus, "Cracks in The Stack: Hidden Vulnerabilities and Licensing Risks in LLM Pre-Training Datasets," 2025 IEEE/ACM International Workshop on Large Language Models for Code (LLM4Code), Ottawa, ON, Canada, 2025, pp. 104-111, doi: 10.1109/LLM4Code66737.2025.00018.

Background - Software Supply Chains

- **Type I**: Dependency-based
- **Type II**: Copy-based reuse
- **Type III**: Knowledge transfer
- **Type IV**: LLM-based reuse (newly emerged)

Motivation

- LLMs are pre-trained on large code corpora (e.g. The Stack v2).
- How good is the curation of training data?
  - Bugs and vulnerabilities
  - License violations
  - Low-quality code
- This may affect quality/usage of LLM-generated code.

Research Goals

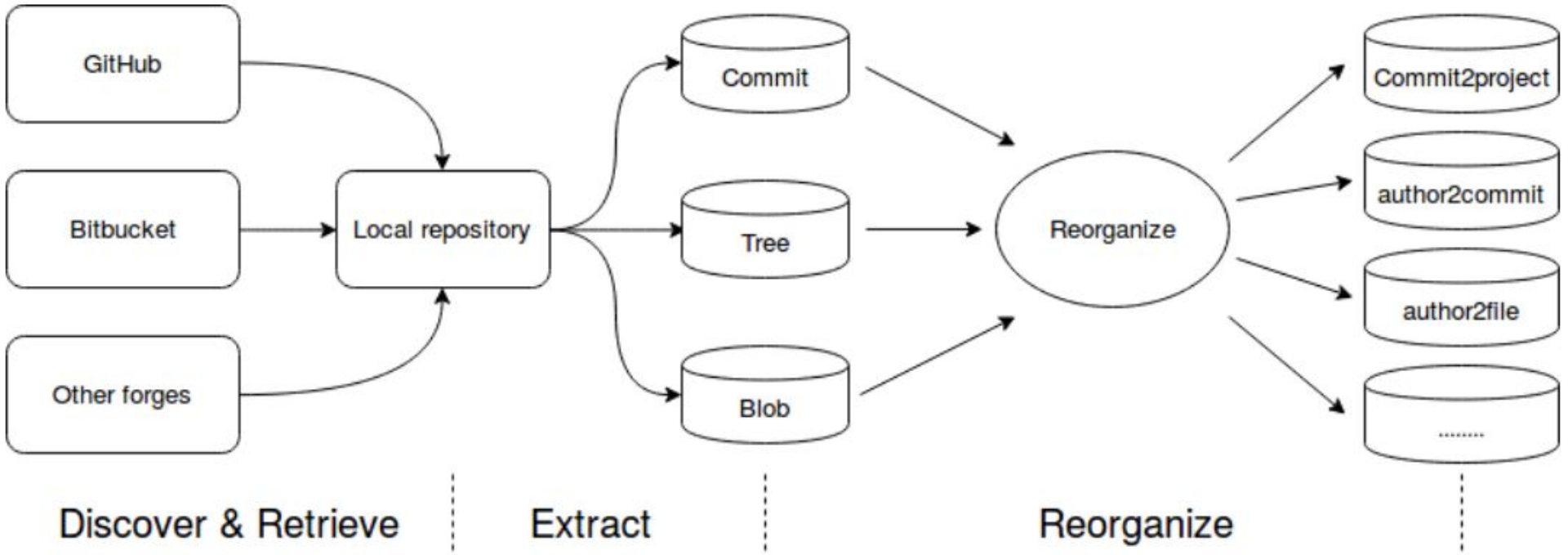
- Evaluate the **quality** of source code in LLM datasets
  - Vulnerable or buggy code
  - Low-use or never-updated code
  - License violations

- Research Questions
  - RQ1: Are there vulnerable or buggy code samples in Stack v2?
  - RQ2: Are there license risks due to reused code?

The Stack v2 Dataset

- The Stack v2 contains over 3B files in 600+ programming and markup languages.
- The Stack serves as a pre-training dataset for open code LLMs.
- In addition to the **full** dataset, the Stack v2 has several deduplicated versions. The-stack-v2-train-**smol**-ids is the most filtered dataset spanning 17 programming languages.

World of Code (WoC)



More information at: https://worldofcode.org/

Methodology Overview

- Extract SHA-1 hashes from Stack v2 blobs.
- Use WoC to:
  - Trace version history for each blob
    - commit-parent commit
    - commit-child commit
  - Identify bug-fixing commits (commit message)
  - Identify code reuse and origin (where each blob was introduced first time)
  - Compare licenses of origin and destination projects for copied blobs

RQ1: Blob Sample

		full		smol	
		count	% (row)	count	% (row)
1	**Total**	**4,553,119**		**680,917**	
2	Missing	115,239	2.53 (1)	16,533	2.42 (1)
3	**Have an old version**	1,622,641	35.63 (1)	287,412	42.20 (1)
	.....				
4	**First version**	2,813,171	61.78 (1)	376,719	55.32 (1)
5	No new version	2,658,805	94.51 (4)	359,380	95.39 (4)
6	**Have a new version**	788,059	**17.30** (1)	69,346	**10.18** (1)
7	Found new versions	1,462,363	-	111,453	-

RQ1: New Version Commit Sample

		full		smol	
		count	% (row)	count	% (row)
1	Commits	835,699		104,782	
2	Blobs	5,068,635		279,652	
3	New versions	5,657,384		307,362	
4	Fix commits	137,091	16.40 (1)	13,628	13.00 (1)
5	Fix blobs	877,811	**17.31** (2)	40,168	**14.36** (2)
6	Fix new versions	935,587	16.53 (3)	41,222	13.41 (3)
7	CVE commits	845	0.61 (4)	83	0.60 (4)
8	CVE blobs	20,765	**2.36** (5)	756	**1.88** (5)
9	CVE new versions	20,561	2.19 (6)	809	1.96 (6)
10	**Distinct CVEs**	**851**		**78**	

RQ1: Key Findings

1. **17.30%** and **10.18%** of blobs in the full and smol datastes, respectively, have newer versions, out of which **17.31%** and **14.36%** are bug fixes.
2. **61.78%** and **55.32%** of blobs are the first version created, out of which **94.51%** and **95.39%** have no newer versions, meaning they were created but never modified, suggesting low quality.
3. There are **19,944** blobs in the clean and deduplicated version of the Stack v2 (smol) that have a newer version were a known security vulnerability is being fixed.
4. In total, **6,947** known CVEs has been found in the smol dataset.

RQ2: Reused Blobs

		full		smol	
		count	% (row)	count	% (row)
1	**Total**	**582,933,549**		**87,175,702**	
2	Reused	90,303,809	15.49 (1)	9,848,987	11.30 (1)
3	Same	29,432,636	32.59 (2)	3,764,702	38.22 (2)
4	Different	60,871,173	67.41 (2)	6,084,285	61.78 (2)

RQ2: License Discrepancies

	Stack v2	WoC	full count	full % (row)	smol count	smol % (row)
1	**Different Origin**		**60,871,173**		**6,084,285**	
2	**Same License**		**38,410,728**	**63.10 (1)**	**4,418,289**	**72.62 (1)**
3	no license	no license	26,604,621	69.26 (2)	3,269,149	73.99 (2)
4	permissive	permissive	11,806,107	30.74 (2)	1,149,140	26.01 (2)
5	**Different License**		**22,460,445**	**36.90 (1)**	**1,665,996**	**27.38 (1)**
6	permissive	no license	10,257,891	45.67 (5)	721,920	43.33 (5)
7	no license	permissive	9,309,959	41.45 (5)	658,085	39.50 (5)
8	no license	restrictive	1,868,500	8.32 (5)	193,358	11.61 (5)
9	permissive	restrictive	1,024,095	4.56 (5)	92,633	5.56 (5)

RQ2: Key Findings

1. **15.49%** and **11.30%** of blobs in the full and smol datasets, respectively, have been reused at least once. Among these, **67.41%** and **61.78%** have origins that were misidentified.
2. **36.90%** and **27.38%** of blobs with misidentified origins have licenses that differ from those identified in the dataset.
3. **12.88%** and **17.17%** of blobs with differing licenses are subject to a restrictive license, presenting a significant risk of noncompliance.

Limitations

- CVE keyword matching may miss some vulnerabilities.
- Not all buggy code is labeled as such.
- Never-modified ≠ definitely unused.
- License assumptions may not apply to all individual files.

Future Work

- Develop **automated curation tools** to:
  - Replace outdated code
  - Remove CVE-prone blobs
  - Filter non-compliant code
- Improve current deduplication approaches

Q&A

Thank You!